CRESST REPORT 823

On the Road to Assessing Deeper Learning: The Status of Smarter Balanced and PARCC Assessment Consortia

January, 2013

Joan Herman & Robert Linn
CRESST/University of California, Los Angeles

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

The work reported herein was supported by grant number 2011-7086 from The William and Flora Hewlett Foundation with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

The authors wish to acknowledge the following individuals and organizations for their contributions to this report: the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers for information access and report review; Eva Baker and Li Cai for their content and technical review; Deborah La Torre Matrundola for monitoring the progress of both consortia, analyzing exemplary prototypes, and manuscript preparation; Ron Dietel for an analysis of content specifications and editorial advice; Tamara Lau for publication design; and Fred Moss for copyediting support.

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the William and Flora Hewlett Foundation.

To cite from this report, please use the following as your APA reference: Herman, J.L. & Linn, R.L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia.* (CRESST Report 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Table of Contents

Executive Summary

Two consortia, the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC), are currently developing comprehensive, technology-based assessment systems to measure students' attainment of the Common Core State Standards (CCSS). The consequences of the consortia assessments, slated for full operation in the 2014/15 school year, will be significant. The assessments themselves and their results will send powerful signals to schools about the meaning of the CCSS and what students know and are able to do. If history is a guide, educators will align curriculum and teaching to what is tested, and what is not assessed largely will be ignored. Those interested in promoting students' deeper learning and development of 21st century skills thus have a large stake in trying to assure that consortium assessments represent these goals.

Funded by the William and Flora Hewlett Foundation, UCLA's National Center for Research on Evaluation, Standards, and Student Testing (CRESST) is monitoring the extent to which the two consortia's assessment development efforts are likely to produce tests that measure and support goals for deeper learning. This report summarizes CRESST findings thus far, describing the evidence-centered design framework guiding assessment development for both Smarter Balanced and PARCC as well as each consortia's plans for system development and validation. This report also provides an initial evaluation of the status of deeper learning represented in both consortia's plans.

Study results indicate that PARCC and Smarter Balanced summative assessments are likely to represent important goals for deeper learning, particularly those related to mastering and being able to apply core academic content and cognitive strategies related to complex thinking, communication, and problem solving. At the same time, the report points to the technical, fiscal, and political challenges that the consortia face in bringing their plans to fruition.

CRESST
National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Full Report

The systems that the Smarter Balanced Assessment Consortium (Smarter Balanced) and Partnership of Readiness for College and Careers (PARCC) are currently developing to assess students' attainment of Common Core State Standards (CCSS) present key leverage points for supporting deeper learning and the development of 21st century competencies essential for students' future success (Pellegrino & Hilton, 2012). These new assessment systems will send a powerful signal to schools about the meaning of the CCSS and what students should know and be able to do for college readiness and success in work. If history is a guide, educators will align curriculum and teaching to what they believe is being assessed and will use assessment results to focus further on teaching and learning to improve performance (Hamilton, Stecher, & Yuan, 2008; Herman, 2010). What is not assessed will largely be ignored.

To the extent that the consortia's assessments call for deeper learning and reflect 21st century competencies, they have the great potential to facilitate teaching and learning that supports these goals. The converse also is true: Absent strong representation in the new assessments, students' deeper learning likely will be compromised. This essential relationship between what is assessed and what is taught provides the context for the CRESST study reported here. Funded by the William and Flora Hewlett Foundation, the project is monitoring the development efforts of both assessment consortia, with particular regard to their attention to deeper learning. Our methodology for defining deeper learning draws on Norman Webb's depth of knowledge (DOK) classification scheme (see Webb, Alt, Ely, & Vesperman, 2005; http://wat.wceruw.org) because it is: 1) commonly used in alignment studies of current state tests, 2) familiar to policymakers and practitioners across the country, and 3) has been used in prior Foundation studies to establish a baseline for current state tests. Webb's system categorizes DOK into the following four levels, essentially:

- DOK1: Recall of a fact, term, concept, or procedure; basic comprehension.
- DOK2: Application of concepts and/or procedures involving some mental processing.
- DOK3: Applications requiring abstract thinking, reasoning, and/or more complex inferences.
- DOK4: Extended analysis or investigation that requires synthesis and analysis across multiple contexts and non-routine applications.

We believe that DOK levels 3 and 4 reflect essential capabilities for 21st century competence that heretofore have been grossly underrepresented in most state tests. In the following sections, we first describe the consortia's current plans for system development and summarize the evidence-centered design framework that guides their work. The framework also undergirds our approach to monitoring the consortia's representation of deeper learning. We then discuss similarities and differences in consortia approaches to implementing the framework and share our analysis of their progress thus far. We end by considering challenges in monitoring and supporting adequate representation of deeper learning in PARCC and Smarter Balanced operational tests.

Assessment System Components

PARCC and Smarter Balanced both aim to create systems of assessment that can serve both formative and summative functions (see Table 1)—although the summative components currently are receiving the lion's share of attention. To serve summative purposes, both consortia are developing measures that combine an on-demand, technology-administered assessment with extended performance tasks in both English language arts (ELA) and mathematics. With Smarter Balanced, both measures will be included as part of an end-of-year assessment, while PARCC will be administering its performance tasks earlier in the spring. Student scores will be aggregated across both contexts and used to characterize student proficiency relative to the CCSS. Scores also are intended to document the extent to which students are on-track to being college and career ready. The two consortia's assessment plans differ in their approach to the on-demand assessment. Smarter Balanced end-of-year, on-demand assessments will utilize computer-adaptive testing (CAT), where complex algorithms are used to customize the items administered to each individual based on his or her ability level, which is inferred from responses to prior items. PARCC's technology-based, end-of-year assessment, in contrast, will use standard, fixed test forms.

PARCC	Smarter Balanced
Diagnostic	Formative practices and tools
Mid-Year benchmark assessment (performance)	Interim assessments
Performance-based assessment	End-of-year performance tasks
End-of-year on-demand test	End-of-year on-demand test (CAT)

Table 1. PARCC and Smarter Balanced Assessment System Components

Furthermore, while both consortia will test students in ELA and mathematics in grades 3-8, they differ in their high school approach. Smarter Balanced plans a summative assessment for students in grade 11 only. PARCC will be testing students in ELA in grades 9-11. However, for high school mathematics, they will use end-of-course tests in Algebra 1, Geometry, and Algebra 2. Both consortia plan to assess every student on a full range of DOK or cognitive complexity to encourage schools to provide opportunities for deeper learning for all students.

On the formative assessment front, PARCC plans to provide teachers and schools with tools they can use throughout the year to diagnose student learning needs relative to the CCSS, and mid-year, benchmark performance assessments intended to provide feedback on students' preparation for the summative, end-of-year performance tasks. Smarter Balanced, in contrast, will provide districts and schools with item banks from which they can construct interim tests to monitor and diagnose student progress and/or serve predictive purposes by mirroring the end-of-year assessment. Smarter Balanced also will make available professional development resources and tools to support teachers' formative assessment practices.

Evidence-Centered Design Framework

Both consortia have adopted Evidence-Centered Design (ECD) as their approach to summative assessment development and validation. Formulated by Robert Mislevy and colleagues (see, for example, Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 1999), ECD starts with the basic premise that assessment is a process of reasoning from evidence to evaluate specific claims about student capability. In essence, student responses to assessment items and tasks provide the evidence for the reasoning process, and psychometric and other validity analyses establish the sufficiency of the evidence for substantiating each claim (see also Pellegrino, Chudowsky, & Glaser, 2001).

ECD is a principled approach that proceeds through a series of interrelated stages to support the close correspondence between the claims and assessment evidence that is used to draw inferences from student scores. As the following describes further, each of these stages represents a critical leverage point in assuring that consortia assessments will well represent the CCSS and deeper learning goals.

The ECD process starts with a clear delineation of the claims that are to be evaluated, and eligible content and skills are closely delimited in a domain model for each claim. The domain model specifies the specific evidence—assessment targets—that can be used to evaluate student status relative to the claim. Item or task models—or specifications—are then developed to guide assessment development. The models provide reusable templates for creating items and tasks aligned with each potential assessment target. These item and task models are then used to generate actual test items and tasks, which, in turn, are subjected to content and bias reviews, pilot tested and field tested, and revised as necessary to refine psychometric quality and validity. At the same time, test blueprints are developed to guide the creation of test forms, which are the collections of items and tasks to which students respond for a given test. The blueprints specify how items and tasks are to be sampled by assessment claim and how targets are to be allocated to test forms—for example, how many items and/or tasks representing specific claims and targets will appear on each test form. The administration of operational test forms then provides additional evidence of psychometric quality and

validity of the tests for their intended purpose(s). Standard setting is a final step in this transparent process.

Figure 1 lays out these stages in the context of the CCSS assessment development. The ECD process here starts with the CCSS themselves. Both PARCC and Smarter Balanced have reorganized the standards into core claims about student competency in ELA and mathematics that their tests are designed to evaluate. Both consortia start with an overall claim about students becoming college and career ready in ELA and mathematics and then subdivide these overall expectations into more specific sub-claims. Tables 2 and 3 summarize PARCC and Smarter Balanced claims for each subject area (We return later to an analysis of these claims.).

Each claim is further defined by specific evidence statements (PARCC) or assessment targets (Smarter Balanced) that the claim encapsulates. These statements or targets, in turn, are operationalized relative to particular standards and/or clusters in the CCSS and specify the DOK or cognitive complexity at which each may be assessed.

In essence, the claims and evidence statements or assessment targets circumscribe the domains to be assessed in terms of content and performance expectations, which then become the targets for developing item and task specifications. The specifications provide guidance and rules that the item writers will follow in developing items that comport with each assessment target or evidence statement. They circumscribe, among other elements, eligible stimulus materials, prompt content and design, response requirements and scoring criteria, accessibility requirements, administration directions and administration conditions (e.g., allowable supports), etc. The ideal item or task specification provides sufficient guidance so that two item writers working independently from the same specification would generate essentially comparable items or tasks for a given assessment target or evidence statement—such that students would be expected to perform similarly on both.

Item writers then use the specifications to generate items and tasks, which in turn are subjected to content and bias reviews as well as pilot testing. Items and tasks which

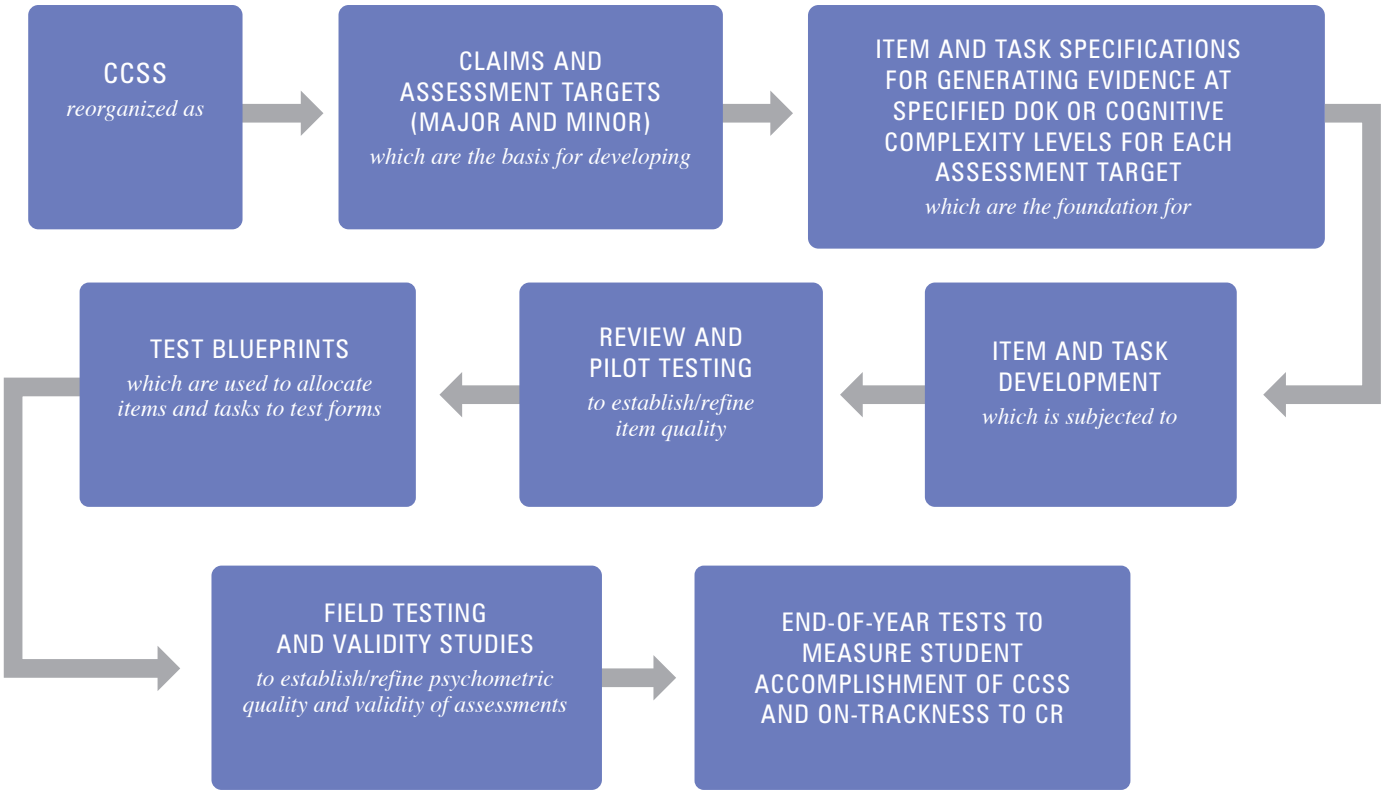

Figure 1. Evidence-Centered Design (ECD) General Approach.

PARCC	Smarter Balanced
1. Reading: Students read and comprehend a range of sufficiently complex texts independently.	1. Reading: Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
2. Writing: Students write effectively when using and/or analyzing sources.	2. Writing: Students can produce effective and well-grounded writing for a range of purposes and audiences.
3. Research: Students build and present knowledge through research and the integration, comparison, and synthesis of ideas.	3. Reading: Students read and comprehend a range of sufficiently complex texts independently.
	4. Research/Inquiry: Students can engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.

Table 2. PARCC and Smarter Balanced Claims for the ELA Summative Assessments

PARCC	Smarter Balanced
1. Major Concepts and Procedures: Students solve problems involving the major content for grade level with connections to practices.	1. Concepts and Procedures: Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
2. Additional and Supporting Concepts and Procedures: Students solve problems involving the additional and supporting content for their grade level with connections to practice.	2. Problem Solving: Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies.
3. Expressing Math Reasoning: Students express mathematical reasoning by constructing mathematical arguments and critiques	3. Communicating Reasoning: Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
4. Modeling Real World Problems: Students solve real world problems engaging particularly in the modeling practice	4. Modeling and Data Analysis: Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
5. Fluency: Students demonstrate fluency in areas set forth in the Standards for Content in grades 3-6.	

Table 3. PARCC and Smarter Balanced Claims for the Mathematics Summative Assessments

survive this process as substantively and psychometrically sound are then assigned to test forms according to blueprints, which provide rules for representing the given claims and assessment targets/evidence statements and for sampling items and tasks. Student performance on annual tests based on the blueprints then provide sound evidence to evaluate the claims and student accomplishment of the CCSS.

The ECD framework thus makes very transparent what is to be assessed and how the CCSS are going to be assessed, and is very different from the "black-box" test development process that has been typical in many state assessments. The black-box process starts with standards and general test blueprints about content coverage, and ends with scores and established proficiency levels, with limited rationale or evidence of the steps in between.

The transparency of the various ECD stages also provides a means for trying to assure that the PARCC and Smarter Balanced assessments will both fully represent the depth and breadth of the CCSS and incorporate deeper learning.

Each stage influences and constrains subsequent ones, and any omissions in prior stages will result in those same alignment gaps in the actual test. For example, if some CCSS are not fully represented in the claims and assessment targets/evidence statements, they will not be included on the test. Similarly, the depth of knowledge or cognitive complexity assessed could look very balanced based on item and task specifications or even within item and task pools that are developed, but if the test blueprint does not adequately sample higher levels of DOK, the test will under-represent them. Analysis of these various documents for PARCC and Smarter Balanced thus represents an important early warning strategy for supporting the content validity of the tests relative to both CCSS and deeper learning goals. In the next sections, we describe CRESST progress in conducting these analyses.

Similarities and Differences in Consortia Approaches

Both PARCC and Smarter Balanced have articulated the major claims about student performance that will be substantiated by their assessments, have laid out general progressions through which they expect teaching, learning, and assessment to develop, and have identified major versus supplementary standards for instruction and assessment. For Smarter Balanced, all the relevant information is contained in a single public document for each subject area, *Content Specifications for English Language Arts and Content Specifications for Mathematics.* Smarter Balanced *Content Specifications* also establish both assessment targets for each of its major claims and the DOK levels at which each target may be assessed. Smarter Balanced *Content Specifications* thus clearly and publically define and circumscribe the domain model to be assessed.

Smarter Balanced issued a request for proposals (RFP) to develop task and item specifications aligned with its *Content Specifications*. Item and task specifications were developed for selected response, constructed and extended response, and technology-enhanced items as well as for performance tasks in ELA and mathematics. Organized by claim, there are separate item or task specifications for each assessment target at each specified DOK level, and each contains at least one task design model and often multiple task models for assessing each level of DOK. Sample items were to be included for each model.

> More recently, PARCC released an RFP for item tryouts and field testing...

The contract has been completed; all available item and task specifications are now posted online (see http://www.smarterbalanced.org/smarter-balanced-assessments). There is a one-to-one correspondence with the *Content Specifications'* assessment targets and DOK requirements, but the task design models are much abbreviated, and sample items have been provided for only a subset of the item and task specifications. As a result, it is difficult to ascertain the actual correspondence in DOK. Smarter Balanced plans to refine its specifications as it develops items and tasks, a contract for item development has been established, and item and task development are currently underway, as is a contract for specifying the test blueprint (see http://www.smarterbalanced.org/smarter-balanced-assessments/ for the preliminary blueprints). Teachers from each Smarter Balanced Governing State were involved last summer in item and task development. In addition, sample performance tasks and items have been released.

In contrast, PARCC's full domain model is still under development. PARCC issued *Content Frameworks* for both ELA and mathematics, which essentially are intended to provide guidance for curriculum and instruction rather than assessment. These documents organize the standards to highlight key advances in learning and understanding that are expected from one grade to the next. The *Frameworks* also provide examples of major within grade dependences among standards and examples of instructional opportunities that connect individual standards and clusters. For mathematics, the *Framework* also presents exemplars for connecting mathematical content and practice standards. It also specifies content emphasis for each grade by identifying mathematics clusters that should be considered "major" versus those that should be considered "additional," or "supporting."

Further keys to the PARCC domain model are found in its Invitations to Negotiate (ITN) and Requests for Proposals (RFPs). The ITN for *Item Development* describes the major claims that PARCC's assessment will address and, as part of its item design framework, provides sample evidence statements that should support each claim and sub-claim (see http://myflorida.com/apps/vbs/vbs_www.ad.view_ad?advertisement_key_num=98159 for contents of the ITN). The document also includes sample item and task specifications, sample tasks, and items of various types and sample test blueprints. This ITN indicates that PARCC will provide successful vendors with a full set of evidence statements, and that vendor development responsibilities will include the item and task specifications as well as the item and tasks themselves (Florida Department of Education, Bureau of Contracts, Grants and Procurement Management Services, 2012). Multiple vendors have been funded for an initial stage, and subsequent funding will be contingent on initial performance. Item and task prototypes have been released. More recently, PARCC released an RFP for item tryouts and field testing, which includes additional blueprint specifications as well as item and task prototypes at various grade levels (see http://www.in.gov/idoa/proc/bids/RFP-13-29/ for contents of the RFP). These collected documents provide evidence for inferring the consortium's current plans for representing deeper learning. In the near future, PARCC plans to create and release a complete form of the specifications in digestible form for the general public.

Representation of Deeper Learning

As noted earlier, to be consistent with prior research documenting the intellectual rigor of current state tests, we are using Norman Webb's DOK methodology to analyze the consortia's representation of deeper learning (Webb et al., 2005). The following analysis starts with an examination and comparison of the major claims guiding each assessment consortium and then moves to an analysis of the explicit or implicit domain models each has established for assessment. For Smarter Balanced, we provide a detailed analysis of the representation of deeper learning in its *Content Specifications* and sample items. For PARCC, we provide an analysis of the initial item design framework and recently released item and task prototypes.

Deeper Learning in PARCC and Smarter Balanced Claims.

Tables 2 and 3, as previously presented, reveal striking similarities between PARCC and Smarter Balanced major claims, as might be expected given that both sets are derived from the same Common Core State Standards. In ELA, both consortia feature students' ability to read and analyze increasingly complex texts, write for a variety of audiences and purposes, and conduct research. While PARCC embeds language use within writing, Smarter Balanced places these capabilities within its claims for both writing and for speaking and listening. PARCC currently does not have speaking and listening claims, because these skills will not be part of PARCC's formal end-of-year assessment. However, PARCC plans future development of claims and assessments for speaking and listening as part of a required, local assessment component that will be part of its system.

In mathematics (see Table 3), both consortia's claims emphasize the application of mathematics knowledge to solve problems, reason mathematically, and communicate one's reasoning. Both also emphasize solving authentic, real world problems. In so doing, they both attend to the integration of mathematical content and practices.

In broad strokes, the major claims established by both consortia appear to comport well with goals for deeper learning. This is particularly true in terms of the content knowledge expectations that students need to master and be able to apply core academic content knowledge and cognitive strategy expectations related to critical thinking, complex problem-solving, and effective communications (at least in some aspects, particularly written communication). Neither collaboration, teamwork, nor metacognition (learning how to learn), which form important elements in many conceptions of deeper learning and 21st century competencies, are explicitly included in any claim. However, collaboration may be incorporated into Smarter Balanced performance tasks, and metacognition may well be required in solving the complex, extended problems that both consortia plan as part of their performance task components.

Deeper Learning in Smarter Balanced Content Specifications.

As noted earlier, Smarter Balanced *Content Specifications* establish both the major claims that the Smarter Balanced summative assessment is expected to evaluate and the assessment targets at specified DOK levels that define each claim. For example, in ELA the assessment targets for Claim 1 include key details, central ideas, word meanings, reasoning and evaluation, analysis of relationships within or across texts, text structures/features, and language use. In mathematics, Claim 1 targets refer to clusters in the CCSS mathematics content standards, while those for Claims 2-4 reference processes related to the mathematics practices (e.g., apply mathematics to solve well-posed problems).

CRESST researchers reviewed the assessment targets within each claim and the DOK levels that are specified for each target. It is of note that there are two or more DOK levels associated with many of the individual assessment targets. Our analysis assumes that the number of assessment targets within each claim provides a gross indicator of the relative emphasis that claim might receive on the end-of-year assessment and that the percentage of targets that are specified at DOK3, and particularly at DOK4, provide an indicator of Smarter Balanced attention to deeper learning goals. Separate analyses were conducted by grade level for ELA (grades 4, 8, and 11) and for mathematics (grades 3-8 and 11).

As Table 4 shows, our analysis of Smarter Balanced *Content Specifications for English Language Arts* reveals an average of 35 content targets per grade. Based on the number of targets specified in each area, Claim 1 (reading) gets the largest emphasis, with 40% of the total targets, followed by Claim 2 (writing) with 29%, then Claim 4 (research/inquiry) with 20%, and finally Claim 3 (speaking and listening) with the remaining 11%.

Overall, DOK2 and DOK3 receive the most emphasis, being designated for 46% and 43% respectively of the total assessment targets. DOK1 is designated for about a third of the targets and DOK4 for a quarter of the targets. It is important to keep in mind that multiple DOK levels can be designated for each target and that fidelity of item development, final blueprint rules, and the nature of Smarter Balanced end-of-year performance tasks, will determine the actual distribution of targets and DOK levels that will be administered to students.

In mathematics, the average number of assessment targets per grade is 29 (Table 5). In general, 38% of these targets are specified within Claim 1 (concepts and procedures), 14% are specified within Claim 2 (problem solving), and 24% each for Claims 3 and 4, which deal respectively with communicating reasoning and using modeling and data analysis to solve real world problems (see Table 5). To the extent that these percentages represent relative emphasis to be accorded each claim on the summative assessment, they presage a Smarter Balanced end-of-year assessment that provides substantial attention to deeper learning goals.

In contrast to ELA, DOK levels for mathematics appear to differ by claim. For Claim 1, virtually all targets are to be assessed at DOK1 and/or DOK2, and essentially no targets are at DOK3 and DOK4. For Claim 2, DOK1 receives relatively less attention, and there is a corresponding increase in emphasis on DOK3, which is specified for

	Overall	Claim 1 Reading	Claim 2 Writing	Claim Speaking & listening	Claim 4 Research/Inquiry
Mean # content targets (%)	35 (100%)	14 (40%)	10 (29%)	4 (11%)	7 (20%)
DOK1	33%	19%	30%	75%	43%
DOK2	46%	55%	47%	50%	33%
DOK3	43%	55%	27%	50%	38%
DOK4	25%	24%	20%	8%	38%

Table 4. Smarter Balanced ELA Content Specification, Grades 4, 8, and 11
Mean Percentage of Content Targets Specified at Each DOK Level

	Overall	Claim 1 Concepts and Procedures	Claim 2 Problem Solving	Claim 3 Communicating Reasoning	Claim 4 Modeling and Data Analysis
Mean # content targets (%)	29 (100%)	11 (38%)	4 (14%)	7 (24%)	7 (24%)
DOK1	46%	82%	50%	0%	29%
DOK2	79%	81%	100%	71%	71%
DOK3	49%	5%	50%	86%	86%
DOK4	21%	0%	0%	43%	43%

Table 5. Smarter Balanced Mathematics Content Specification, Grades 3-8 and 11 Mean Percentage of Content Targets Specified at Each DOK Level

half the targets. For Claims 3 and 4, DOK3 gets the most attention and DOK4 is specified for more than 40% of the targets. As with ELA, the assessment blueprint and nature of Smarter Balanced end-of-year performance tasks will determine the actual representation of various DOK levels in the assessment of any individual student.

Deeper Learning in Smarter Balanced Performance Task Specifications and Sample Items.
As mentioned, CRESST did not conduct a close analysis of Smarter Balanced performance and item specifications because of limitations of the evidence. However, a review of the task specifications and public release sample items suggest that Smarter Balanced performance tasks will be aligned with important goals for deeper learning and will reflect DOK4. The general specifications require that tasks be based in authentic, life-like scenarios and ask students to explore, respond to, and synthesize a variety of stimuli (texts, videos, math applications), which then are the basis for extended written or oral products.

Figures 2 and 3 show examples of Smarter Balanced performance tasks in ELA and mathematics (see http://www.smarterbalanced.org/sample-items-and-performance-tasks/ for available sample items and performance tasks). In both cases, the samples involve a whole class activity as well as multiple student tasks. The whole class activity serves the purpose of introducing students to a familiar scenario, enabling students to share their prior knowledge, and in the case of mathematics performance tasks, provide original data for students to use. Once students have been oriented towards the scenario, students engage in successive activities that build to DOK4. Students are required to analyze multiple stimuli (i.e., web sites about nuclear power, class data, and a cost chart for different field trips), synthesize and evaluate information by completing extended, constructed response items, and compose multi-paragraph arguments for a specific audience (i.e., congressman, teacher) using evidence from multiple stimuli.

Based on existing examples, technology and language potentially add construct irrelevant demands to some tasks. However, Smarter Balanced is developing an accessibility framework, and will conduct cognitive lab and accessibility and bias reviews that should serve to identify and reduce such challenges.

Deeper ELA Learning in PARCC Request for Item Development and Task Prototypes.
As noted earlier, PARCC's request for item development included the Consortium's item design framework. In it, PARCC establishes the claims and sub-claims its assessments will use to measure ELA and mathematics performance. As described earlier, the overall claim for each subject area is that students are on-track to college and career readiness, with the specific claims and major sub-claims articulating the critical competencies in each subject area. Evidence statements for each claim and sub-claim provide a third element of the framework. The ITN provides sample evidence statements—corresponding to Smarter Balanced assessment targets; each of these are to reflect individual standards and/or clusters of standards. The ITN notes that contractors will be provided with a final and complete list of evidence statements, with expected levels of cognitive complexity specified for each. Sample

Classroom Activity (20 minutes): Using stimuli such as a chart and photos, the teacher prepares students for Part 1 of the assessment by leading students in a discussion of the use of nuclear power. Through discussion:

- Students share prior knowledge about nuclear power.
- Students discuss the use and controversies involving nuclear power.

Part 1 (50 minutes): Students complete reading and pre-writing activities in which they:

- Read and take notes on a series of Internet sources about the pros and cons of nuclear power.
- Respond to two constructed-response questions that ask students to analyze and evaluate the credibility of the arguments in favor and in opposition to nuclear power.

Part 2 (70 minutes): Students individually compose a full-length, argumentative report for their congressperson in which they use textual evidence to justify the position they take pro or con on whether a nuclear power plant should be built in their state.

Scoreable Products: Responses to the constructed-response questions and the report.

Figure 2. Smarter Balanced Sample Performance Assessment Task, 11th Grade ELA: Nuclear Power: Friend or Foe? Adapted from Grade 11 Performance Task, by Smarter Balanced Assessment Consortium, 2012, Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/nuclear.pdf. Reprinted with permission.

item types as well as item and sample blueprints provide additional elements of the framework. We examine next how these elements comport with deeper learning goals.

The ITN's articulation of major claims and sub-claims encapsulate important goals for deeper learning:

1) **Reading complex text:** Students read and comprehend a range of sufficiently complex texts independently.

   a) **Sub-claim 1.1:** Close analytic reading (CAR) in literature.
   b) **Sub-claim 1.2:** Close analytic reading (CAR) in informational texts (history/ss, science, technical subjects).
   c) **Sub-claim 1.3:** Vocabulary interpretation and use.

2) **Writing:** Students write effectively when using and/or analyzing sources.
   a) **Sub-claim 2.1:** Written expression: Students produce clear and coherent writing in which development, organization, and style are appropriate to the task, purpose, and audience.
   b) **Sub-claim 2.2:** Conventions and knowledge of language: Students demonstrate knowledge of conventions and other important elements of language.

3) **Research** (major claim for grades 6-11/Sub-claim 3.1 for grades 3-5): Students build and present knowledge through research and the integration, comparison, and synthesis of ideas.

Sample ELA item and task types described by the ITN similarly appear to represent high degrees of intellectual rigor. For example, draft performance task generation models focus on "engaging with literature/literary analysis," and "research simulation." Each model requires students to read multiple texts, construct responses to analysis and synthesis questions about the texts, and produce extended essays that also call for students to analyze and synthesize the materials read. The preliminary task blueprint calls for sessions over two days for each performance task. Contractors are required to develop three task generation models for each task type.

Further, PARCC recently released ELA item and task prototypes at multiple grades spanning elementary to

Classroom Activity: Teacher introduces students to the topic and activates prior knowledge of planning field trips by:

- Leading students in a whole class discussion about where they have previously been on field trips, with their school or youth group.
- Creating a chart showing the class's preferences by having students' first list and then vote on the places they would most like to go on a field trip, followed by whole class discussion on the top two or three choices.

Student Task: Individual students:

- Recommend where their class should go on a field trip, using their analysis of the class vote to justify their response.
- Determine the per-student cost of going on a field trip to three different locations, based on a given chart showing the distance, bus charges, and entrance fees for each option
- Use information from the given cost chart to evaluate a hypothetical student's recommendation about going to the zoo.
- Write a short note to their teacher recommending and justifying which field trip the class should take, based on an analysis of all available information.

Scoreable Products: Answers to the four constructed-response tasks.

Figure 3. Smarter Balanced Sample Performance Assessment Task, 6th Grade Mathematics: Taking a Field Trip. Adapted from Grade 6 Performance Task, by Smarter Balanced Assessment Consortium, 2012, Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/fieldtrip.pdf. Reprinted with permission.

high school (see http://www.parcconline.org/samples/item-task-prototypes). Figure 4 summarizes a grade 7 performance task about Amelia Earhart. Like the Smarter Balanced example, the task subsumes three distinct activity components that build upon one another to a final DOK4 task, but initial components function differently. PARCC introduces and orients students to the topic of Earhart by having students individually read and write a summary about a single text. Students then complete and respond to additional reading, which also functions as a pre-writing activity. Students are asked to analyze claims about Earhart and her navigator using evidence from a second text. Finally, after reading and responding to a third text, students write a multi-paragraph essay in which they use textual evidence to analyze the strength of the arguments made about Earhart's bravery in at least two of the texts.

Deeper Mathematics Learning in PARCC Request for Item Development and Task Prototypes.

As in ELA, the mathematics claims presented in the ITN appear to align with goals for deeper learning. The claims specifically integrate both mathematical content and practices:

1) **Major content with connections to practices:** The student solves problems involving the major content for her grade/course with connections to standards for mathematics practices.

2) **Additional and supporting content with connections to practices:** The student solves problems involving additional and supporting content for her grade/course with connections to standards for mathematics practices.

3) **Highlight mathematical practices MP.3, MP.6 with connections to content:** Expressing mathematical reasoning. The student expresses grade/course level appropriate mathematical reasoning by constructing viable arguments, critiquing the reasoning of others and/or attending to precision when making mathematical statements.

4) **Highlight practice MP.4 with connections to content:** Modeling/application. The student solves real world problems with a degree of

Summary Essay: Using textual evidence from the Biography of Amelia Earhart, students write an essay to summarize and explain the challenges Amelia Earhart faced throughout her life.

Reading/Pre-Writing: After reading Earhart's Final Resting Place Believed Found, students:

- Use textual evidence to determine which of three given claims about Earhart and her navigator, Noonan, is the most relevant to the reading.
- Select two facts from the text to support the claim selected.

Analytical Essay: Students:

- Read a third text called Amelia Earhart's Life and Disappearance.
- Analyze the evidence presented in all three texts concerning Amelia Earhart's bravery.
- Write an essay, using textual evidence, analyzing the strength of the arguments presented about Amelia Earhart's bravery in at least two of the texts.

Scoreable Products: Summary essay, analytical essay, and answers to the comprehension questions.

Figure 4. PARCC Performance-Based Research Simulation Task Prototype, 7th Grade ELA: Amelia Earhart Adapted from Grade 7 – ELA/Literacy, by Partnership for Assessment of Readiness for College and Careers (PARCC), 2012, Retrieved http://www.parcconline.org/samples/english-language-artsliteracy/grade-7-elaliteracy. Reprinted with permission.

difficulty appropriate to the grade/course by applying knowledge and skills articulated in standards, engaging particularly the modeling standard and, where helpful, making sense of problems and persevering to solve them (MP.1) by reasoning abstractly and quantitatively, using appropriate tools, making sense of structure, etc.

5) **Fluency in applicable grades (Grades 3-6):** Student demonstrates fluency as set for the relevant grade level standards.

The ITN defines three broad classes of item types for addressing these claims, each of which encompasses innovative item formats:

- **Type I:** Machine scoreable, balance of conceptual understanding, procedural knowledge, and brief applications; includes innovative item types and technology enhanced formats to get a deeper level of understanding.

- **Type II:** Extended response and/or innovative technology-enhanced formats that call for written arguments, justification, critique, and/or evaluation of precision; incorporates practice forward and integrative tasks.

- **Type III:** Generate evidence for Claim 4, solving complex, novel problems; scenario based; tasks are practice-forward, highlighting modeling among other mathematics practices.

Preliminary blueprints for both the end-of-year on-demand assessment and performance tasks are specified in terms of each of these item types to operationalize sample test blueprints. According to the ITN, the on-demand assessment will be comprised of Type I items and will address the first two claims, involving using major or additional/supplementary content knowledge to solve problems, and the fluency claim for grades 3-6. The performance tasks will be used to address Claim 3 (mathematical reasoning and communication) and Claim 4 (modeling to solve problems). Tasks will feature real world scenarios, integrative tasks that require the synthesis of multiple content standards, and "practice forward" tasks in which application of a mathematical practice is essential—without it, the problem cannot be solved.

Part A: Students analyze data from an experiment involving the effect on the water level of adding golf balls to a glass of water in which they:

- Explore approximately linear relationships by identifying the average rate of change.
- Use a symbolic representation to model the relationship.

Part B: Students suggest modifications to the experiment to increase the rate of change:

Part C: Students interpret linear functions using both parameters by examining how results change when a glass with a smaller radius is used by:

- Explaining how the y-intercepts of two graphs will be different.
- Explaining how the rate of change differs between two experiments.
- Using a table, equation, or other representation to justify how many golf balls should be used in each experiment.

Scoreable Products: Answers to the five constructed-response tasks

Figure 5. PARCC Performance-Based Mathematics Task Prototype, High School: Golf Balls in Water Adapted from The Mathematics Common Core Toolbox, by The Charles A. Dana Center at the University of Texas at Austin and Agile Mind, Inc., 2012, Retrieved from http://ccsstoolbox.agilemind.com/parcc/about_highschool_3834.html. Reprinted with permission.

The ITN further specifies that all end-of-year tests will address each of the mathematical practices. Further, in determining students' final status with regard to being on track to college and career readiness, the ITN estimates that the performance tasks will constitute approximately 40% of the final score.

Figure 5 shows a PARCC prototype performance task for high school mathematics. It exemplifies DOK4 through a multipart investigation of linear relationships using an experiment involving the effect on the water level of adding golf balls to a glass of water. In Part 1 of the task students are provided with a graphic of the problem (a picture of a glass of water with two golf balls) and a table showing the results of 5 trials with different numbers of balls; students are asked to write an equation showing the relationship between the number of balls and the water level. Part 2 builds on this case by having students suggest a modification to the experiment to explore rates of change further. In Part 3, students analyze the results of the modification and provide evidence (e.g., table, equation, etc.) for their conclusions.

Consortia Intents Compared to Current State Assessment Practice.

It appears that the consortia are moving testing practice forward substantially in representing deeper learning, but the nature of available data make it difficult to determine precisely the extent of the change. First of all, data on consortia plans for assessing deeper learning discussed in this report are based on the full domain of potential assessment targets from which items and tasks will be selected for testing. Given time constraints for test administration, it will not be possible for a single test form to include all targets. Secondly, existing comparison data on representation of deeper learning in current state tests comes from the analysis of actual tests—not from detailed specifications for current tests, because they do not generally exist. Available data for both consortia and current tests, thus, are not directly comparable.

With this important caveat in mind, we look to a recent RAND study (Yuan & Le, 2012) to gauge current state practice and compare it to consortia plans. The RAND researchers analyzed the DOK of released test items and

tasks in reading, writing, and mathematics from 17 leading states, which prior research showed as the most likely to have cognitively demanding tests that addressed deeper learning. These were the states that used open-ended and writing tasks, in contrast to most states that relied solely on selected response and short constructed response items in their tests. RAND's analysis assumes that released items provide an indicator of current assessment practice, although the degree to which released items adequately reflect the full operational tests used by states is unknown.

Using Norman Webb's DOK scheme, the analysis of mathematics items revealed the cognitive rigor of all selected response items across states at or below DOK2, with the majority at DOK1. States' open-ended items fared slightly better; on average, 88% were at DOK1 and DOK2, and 11% at DOK3. Based on these estimates, Smarter Balanced specifications portend a dramatic upward shift in intellectual rigor and toward deeper learning.

A higher degree of rigor was observed in reading than in mathematics, with selected response items ranging from DOK1 to DOK3, and with a mean of 14% of items at DOK3. For open-ended reading items, 49% and 11% respectively of state items were at DOK3 and DOK4. For the 12 states that had a separate test of writing, a mean of 47% of the selected response items were at DOK1 and DOK2, and 33% at DOK3. For the eight states that had open-ended writing samples, 47% and 44% respectively were at DOK3 and DOK4.

RAND estimates of available state assessment items and tasks in ELA cannot be directly compared with Smarter Balanced content specifications because the latter combine reading and writing, while the RAND data do not allow easy summary across the two (Yuan & Le, 2012). Further, even among the leading states in the RAND study, many only assessed reading and less than half included extended writing tasks. The latter is worth underscoring in that the RAND data indicate that open-ended, extended tasks are a key to achieving deeper learning at DOK4.

RAND's summary statistics on United States (US) students' exposure to deeper learning also provide an instructive comparison (Yuan & Le, 2012). The RAND researchers defined a state as holding students accountable for deeper learning if at least 5% of available selected response items and/or one open-ended task reflected DOK4. The more stringent criterion involved both item types, while the less stringent criterion focused only on open-ended tasks. Assuming that none of the state assessments in non-leading states addressed deeper learning and using 2009-2010 student enrollment data, Yuan & Le (2012) estimated that 0% of students in the US were assessed on deeper learning in mathematics through state tests, 16% percent of students were assessed on deeper learning in reading, and 2-3% were assessed on deeper learning in writing. Overall, 3-10% of US elementary and secondary students were assessed on deeper learning on at least one state assessment.

The situation will be very different if Smarter Balanced assessments reflect its content specifications and PARCC follows its current plans. By RAND's metric, fully 100% of students in tested grades using consortia tests will be held accountable for deeper learning.

> By RAND's metric, fully 100% of students in tested grades using consortia tests will be held accountable for deeper learning.

Discussion and Conclusion

CRESST analysis thus far indicates that both PARCC and Smarter Balanced summative assessments and students' proficiency classifications based on the assessments will represent many goals for deeper learning, particularly those related to mastering and being able to apply core academic content and cognitive strategies related to complex thinking, communication, and problem solving. We find such representation in the claims that both consortia have established as the foundation for their assessments as well, as in our analyses of underlying content specifications for

the tests, and of sample test and item specifications, sample items and performance tasks, and sample test blueprints.

Our analysis of Smarter Balanced content specifications, for example, found that as many as 49% of its mathematics assessment targets may be assessed at DOK3 and as many as 21% at DOK4; for ELA the corresponding figures are 43% at DOK3 and 25% at DOK4. These expectations reflect a dramatic increase in intellectual rigor relative to current state assessments. RAND's recent study (Yuan & Le, 2012), for example, found no representation of higher levels of depth of knowledge in state mathematics tests.

> "...key questions remain about how well these intentions will be realized."

Our analysis thus far reflects the intentions of both consortia. However, at this early stage of test development, key questions remain about how well these intentions will be realized. As we noted earlier, for example, Smarter Balanced specified multiple levels of depth of knowledge for many of its assessment targets in ELA and for most in mathematics. Final test blueprints will determine how well higher versus lower level targets are represented on the actual tests, and validity studies—such as cognitive labs planned by Smarter Balanced—will reveal how well selected items and tasks reflect their intended levels. How well they capture DOK4, which represents the heart of goals for deeper learning, will depend on the nature of the performance tasks that are included as part of each consortium's system.

The performance tasks themselves represent a potential danger point. Some of the Chief State School Officers in the Smarter Balanced Governing States, for example, pushed back on Smarter Balanced initial plans for multiple performance tasks over several days because of time demands and the cost burdens of scoring. Smarter Balanced now plans to reduce the time requirements for its summative assessment while still being able to provide a claim-level report for every student.

Given the stresses on budgets in many states, there is danger that states in both consortia may try to cut back even further and perhaps even omit the performance tasks to save costs. In addition to costs, extended performance tasks also offer a challenge in assuring the comparability of scores from one year to the next. Without comparable or equitable assessments from one year to the next, states' ability to monitor trends and evaluate performance may be compromised. Responding to the challenge may well require innovation in performance task design, scoring, and equating methods.

Both consortia have been optimistic about the promise of automated constructed-response and performance task scoring and have incorporated that optimism into their cost estimates for the summative assessment. Both are estimating summative testing costs at roughly $20 per student for both subject areas. In the absence of promised breakthroughs, those costs will escalate, there will be enormous demands on teachers and/or others for human scoring, and the feasibility of timely assessment results may be compromised. For example, Smarter Balanced has promised end-of-year results, including those from performance tasks, within two weeks of the end of its spring testing window, and PARCC has committed to quick turnaround times as well. Meeting these expectations also will require innovation in scoring services.

Both consortia have put forth extensive requirements and are convening special advisory groups to support the accessibility of their assessments for students with disabilities and for English language learners. Technology-based assessment provides new, built-in opportunities for customizing an assessment to individual accessibility needs, for example through read-aloud, glossary, and other language supports; text magnification; and other options that can be individually activated. At the same time, while built-in accommodations may be easier to accomplish, there will still be the validity challenge of establishing the comparability of accommodated and non-accommodated versions of the test. Similarly, and particularly in the short run, while technology-based assessment offers many new opportunities, care will need to be taken that the

technology manipulation required by the assessment does not unintentionally add construct-irrelevant barriers for some students, particularly those with less access and less facility with technology. In addition, performance tasks typically include substantial reading or other linguistic demands that can provide unintended obstacles to English learners and those with low reading ability being able to show what they know. Developers will need to be creative in designing tasks, items, and accessibility mechanisms that minimize construct-irrelevant, unintended demands.

The increased intellectual rigor—DOK level—that both consortia's assessments are intended to embody is both a tremendous strength and a potential challenge to implementation. As noted earlier, if initial intentions are realized, consortia assessments will address much deeper levels of knowledge, application, communication, and problem solving than do current state assessments. On the other hand, initial results are likely to provide a shock to the public and to teachers' usual instructional practice. Schools' and teachers' transition to the CCSS and the availability of resources to support that transition will make a tremendous difference in how well the new assessments are accepted and/or whether there is additional pushback to them. Both consortia are helping their states plan for transition, even as that responsibility is well beyond their charge and resources.

As one important aspect of the transition, we would argue that it is important to keep schools focused on the claims and how they can incorporate the individual assessment targets or evidence statements, rather than the individual targets in isolation. The latter are intended to build to the claims and are more than any teacher or their students can reasonably keep in focus or on track. For example, Smarter Balanced content specifications include a relatively large number of assessment targets for each grade—on average 29 targets in mathematics and 35 targets in ELA. The claims, in contrast, reflect a reasonable number of major learning goals and represent the broad competencies that students need for college and career readiness. History suggests that focusing on discrete, individual standards is not the way to develop deeper learning, yet this is the strategy that states, districts, schools, and teachers have typically followed. They will need support to pursue pathways that are more coherent for deeper learning.

The transparency of subsequent steps of the test development and validation process represents yet another possible challenge to assuring the representation of deeper learning in the consortia's efforts. Smarter Balanced has been very transparent in posting all of its plans and the results of its contracts. Yet, because its computer adaptive testing approach essentially individualizes test items to every student, it may be difficult to ascertain how well deeper learning is represented for every student or overall. The test blueprint will provide rules for item selection and presumably, those rules will include those for representing higher levels of depth of knowledge, but this is yet to be seen. There also will be a challenge in clearly communicating to students, teachers, and parents about what has been assessed. More complete analysis of PARCC development targets and processes will be possible in the near future when its plans are complete and made public.

It is clear that both consortia have very ambitious agendas and timelines. There is no doubt that they both have made enormous progress. We will continue to monitor that progress and continue to assess their fidelity to the rigorous demands of the CCSS and to goals for deeper learning. ■

References

Florida Department of Education, Bureau of Contracts, Grants and Procurement Management Services. (2012). *Invitation to negotiate: PARCC item development*. Tallahassee, FL: Author.

Hamilton, L., Stecher, B., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions.* Washington, D.C.: Center on Education Policy.

Herman, J. L. (2010). Impact of Assessment on Classroom Practice. In E. L. Baker, B. McGaw, & P. Peterson (Eds.), *International Encyclopedia of Education* (pp. 506-511). Oxford: Elsevier Limited.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6-20.

Mislevy, R., Steinberg L., & Almond, R. (1999). *Evidence-centered assessment design.* Princeton, NJ: ETS.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington, DC: The National Academies Press.

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). Web alignment tool (WAT): Training manual 1.1. Retrieved June 2, 2012 from http://www.wcer.wisc.edu/WAT/Training%20Manual%202.1%20Draft%20091205.doc

Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests* (WR-967-WFHF). Santa Monica, CA: RAND.